

本周周报

解聪

11/11/2013 – 11/17/2013

本周工作

之前一个月的工作针对企业数据以及淘宝数据进行案例的寻找和系统的实现。这周主要在之前标签项目进行了总结与下一步方向的思考：

1. 相似性的计算

之前一直在找计算人群的购买行为相似的方法，比如先后使用了欧式距离，皮尔森相关系数和 CCA 等方法。问题是这些方法的效果都不是非常满意。比如 CCA 的结果计算出来的人群相似性始终在 0.5 以上（最大 1.0）。而欧式距离均等考虑了所有购买类目，隐藏了关键信息。他们都无法发现局部关联信息。

而这些相似性/距离的算法又影响了投影/聚类算法。比如 MDS，SOM 等，所以之前的投影效果不好，一部分也是受相似性算法的影响。

是不是可以说自己定义人群购买类目的相似性算法来解决这个问题？参考论文（Distance metric learning, with application to clustering with side-information 这篇论文）

对于人群 x 和人群 y ，使用 $1 \times n$ 的向量 X, Y 来代表其不同购买类目上的分布。

如果使用普通的欧式距离，人群 AB 的购买行为的距离的定义为：

$$d(x, y) = \sqrt{(X - Y)^T * (X - Y)}.$$

购买行为的相似性 S 则可以定义为

$$S(x, y) = 1 - d(x, y)$$

现在将距离尺度的定义拓展为：

$$d(x, y) = \sqrt{(X - Y)^T * A * (X - Y)}$$

A 是 $n \times n$ 的矩阵，当 A 为单位阵 E 时，上式为欧式距离。

我们可以通过用户的交互训练得到距离矩阵 A 的最优值。（参考 UTOPIAN: User-Driven Topic Modeling Based on Interactive Nonnegative Matrix Factorization 这篇论文）

1). 通过指定训练集。比如用户指定人群 x_i 和 x_j 是相似的，继而通过这些相似的人群训练出矩阵 A 。即求 A ，使得：

$$\min \sum ||x_i - x_j||_A^2$$

求解可以利用 Distance metric learning 这篇论文里提出的方法。

2). 通过直接限定 A 的值得到用户想要的结果。用户通过 A' 矩阵输入自己想要的矩阵特定的值。即求 A 使得：

$$\min \sum ||x_i - x_j||_A^2 + ||A - A'||_F^2$$

在保证 1)的情况下，使得 A 和用户指定的 A' 尽量接近。

1)和 2) 都可以通过用户交互的方式输入。比如通过对像素图的点击，划选等。

2. 不确定性的分析

另外一方面，先前的工作可视化效果还是略有欠缺。图像模式的出现的过于散乱，不太方便用户找到图像模式。问题有：

1. 我们在处理数据时，将数值属性也变为了类别属性（比如年龄分为了 18-20,20-25,25-30 等，每一段作为一类）。但事实上这样的分法准不准确并不清楚，他可能将不同行为的人合并或将相同行为的人拆开。
2. 不同标签之间是有关联的，存在一些信息冗余。
3. 分类的顺序很关键，比如先按年龄分，后按性别分；还是先按性别分，后按年龄分。两者对用户交互体验影响很大。

对于 2, 3 两个问题有必要在选择属性前先计算一下分割出来的用户群在购买类目分布的信息熵的大小。熵越大代表越混乱，所以应当选取人群分割后所有人群的熵总和最小的属性先分。

其次，后续划分的情况是取决于上一次的划分，计算熵的时候要考虑到相对熵等因素。

对于第一个问题，同样如果分割后人群在商品类目上的分布的熵很大，说明当前的划分的很杂乱，不能体现数据本身的模式。因此我们要寻找一种分法，使得：

$$\min \sum H(X_i)^2$$

X_i 是划分后的各人群在商品类目上的概率密度函数。怎么求解目前还不太清楚。

下周工作

求解本周提出的两个问题，第一个相似性的问题已经有现成方案。第二个熵的求解还需要继续探索。